PREDICTIVE MAINTENANCE OF ESSENTIAL MACHINERY: A DATA-DRIVEN APPROACH FOR OIL AND GAS INDUSTRY RESILIENCE

Team 10
Chia Wei Kit, Samuel
Dhruval Kenal Kothari
Lim En Ning
Riley Ang Xile
Tan Zheng Da Jared

RENAISSANCE ENGINEERING PROGRAMME
NANYANG TECHNOLOGICAL UNIVERSITY

Year 2023/24

Executive Summary

The cost of downtime in the oil and gas industry has more than doubled from 2021 to 2022. With the hourly cost of downtime soaring to nearly half a million dollars in the oil and gas sector in 2022, unplanned downtime cost Fortune Global 500 companies 11% of their annual revenue, amounting to $1.5 trillion, which represents a 70% increase in losses compared the previous two years [1]. Aramco produces 12.8 million barrels of oil equivalent per day [2] and the financial implications of downtime are monumental.

The objective of this project is to address the significant challenge of unplanned equipment downtimes in the oil and gas industry. This project seeks to develop a predictive maintenance framework using the Turbofan Engine Degradation Simulation Data Set [3] to accurately predict the Remaining Useful Life (RUL) of critical machinery. By predicting equipment failures before they happened, this project seeks to minimise unplanned downtimes, reduce maintenance costs, and improve overall safety in Aramco's operations.

The methodology for this project starts with an analysis on the dataset and extends into feature engineering and selection processes where the raw data is transformed into a refined set of features through autocorrelation assessment, temporal feature creation, and irrelevant feature elimination. Machine learning models, both regression and classification types, are then applied to predict the Remaining Useful Life (RUL) of machinery and categorise equipment based on risk levels. The models used range from Linear and Elastic Net Regression to more complex algorithms like Random Forest and XGBoost. Finally the models are evaluated using metrics that are tailored to both regression and classification tasks, like Mean Absolute Error (MAE), Mean Squared Error (MSE), and the Area Under the Receiver Operating Characteristic Curve (ROC AUC), ensuring the selection of models that best predict machinery failures and optimise maintenance strategies. Among various regression models, XGBoost Regressor demonstrated the highest efficiency in predicting the RUL of machinery. However, the model exhibited a degree of bias in estimating high and low RUL values, indicating room for improvement in its predictive consistency. The XGBoost Classifier was effective in distinguishing 'at risk' and 'not at risk' machinery. It also has the highest ROC AUC scores. Despite a slight decrease in performance metrics after hyperparameter tuning, the model showed enhanced generalisation capabilities, making it suitable for real-world applications. An economic analysis was conducted to assess the cost-benefit aspects of predictive maintenance and an optimal threshold for classifying 'at risk' machinery was recommended.

A two-pronged approach is recommended for Aramco to enhance predictive maintenance. Using a regression model for critical and high-value equipment and using a classifier model for less critical components.

Table of Content

1 Introduction

This cost of unplanned downtime is skyrocketing. In the oil and gas industry, the cost of one hour of downtime has more than doubled from 2021 to 2022 to nearly $500,000 [1]. This contributes to an overall loss of $1.5 trillion annually for Fortune Global 500 companies, and the need to adopt a predictive maintenance approach is increasingly important [4]. The implementation of predictive maintenance and leveraging data analytics and real-time monitoring through sensors and IoT can help identify when machines will fail. Machines can then be maintained before failure thus significantly reducing unplanned downtime and its associated costs [4].

For Aramco, any unplanned downtime can lead to staggering financial losses, and potential safety incidents. Furthermore, this trend of increasing cost in unplanned downtime would impact Aramco's profitability and long-term sustainability severely because it handles 12.8 million barrels of oil equivalent per day [2].

To address this challenge, our project aims to propose a pivot in strategy towards predictive maintenance. By leveraging data analytics to anticipate machinery failures before they occur. Using the Turbofan Engine Degradation Simulation Data Set, our objective is to develop a predictive maintenance framework to accurately predict the RUL of machinery. This initiative is expected to drastically reduce unplanned downtimes and cut the maintenance costs of Aramco's operations. Through the use of feature engineering and regression and classification models like XGBoost and Random Forest, we aim to show that raw operational data can be transformed into actionable insights. Through the key findings of this project, we hope to improve the predictive capabilities at Aramco and minimise the cost of unplanned downtimes.

2 Business Problem

Saudi Aramco is a global leader in the oil and gas industry. It is currently facing a growing need to improve the maintenance of its machinery due to rising costs of unplanned downtime. This oil and gas industry is inherently capital-intensive and operations are heavily reliant on the efficiency and reliability of machinery [5]. Unplanned downtimes caused by machine failures not only lead to significant financial losses, it also poses a risk to safety.

The concept of maintenance in most industries revolves around 3 main strategies: reactive, preventive, and predictive maintenance. Each strategy addresses the unavoidable point of failure that every machine will reach in different ways. Reactive maintenance uses a "wait and react" philosophy, where maintenance actions are only performed after a device has failed. While this is acceptable for non-critical appliances, it is extremely costly and inefficient in the industry setting

as it can halt production and cause significant financial losses. Preventive maintenance avoids the drawbacks of reactive maintenance through regular checks and maintenance actions before a machine reaches its point of failure. Although this prevents sudden unexpected failures in an industry setting, performing maintenance too early can lead to wasted RUL in a machine's life, thus not a cost-effective solution. Predictive maintenance overcomes the efficiency concerns of preventive maintenance by accurately forecasting when a machine will fail by predicting its RUL. This approach solves the business problem by reducing downtime and preventing halts in production but also maximising a machine's lifetime to reduce maintenance costs.

There has been an industry shift from a reactive maintenance approach towards a proactive approach which utilises data analytics to predict machinery failures before they occur [6]. Using the suite of sensors and IoT that Aramco has [7] enabled monitoring of equipment health in real-time to facilitate the process of maintaining the equipment based on the current condition of the equipment rather than predetermined schedules or when failure occurs.

The NASA Turbofan Engine Degradation Simulation Data Set is relevant to this project as it allows for exploration and modelling of engine degradation patterns, making it useful for developing predictive maintenance models. Utilising this dataset can help in developing a framework for predicting RUL of critical and non critical machinery, thus enabling Aramco to implement timing maintenance and avoid unplanned outages and production disruptions as well as reduce maintenance costs.

3 Literature Review

3.1 Overview

Predictive maintenance (PdM) strategies have become indispensable in modern industrial settings, offering a proactive approach to equipment management by detecting anomalies and predicting potential failures before they occur. Predictive maintenance techniques use advanced sensor technologies and data analytics to monitor equipment health in real-time, allowing for planned maintenance interventions based on the equipment's actual condition rather than fixed schedules [8]. By preemptively addressing issues, PdM enhances equipment uptime, reduces maintenance costs, and optimises spare part inventory management.

3.2 Comparison with Traditional Maintenance Approaches

A common strategy used in the past is corrective maintenance, also known as "run-to-failure" which involves replacing a part only when it is damaged and equipment is unable to work without any assistance or intervention [9]. However, this leads to high costs and longer

turnaround times where certain machinery is not available to replace immediately. Hence, resulting in loss of machine efficiency and longer down times. On the other hand, predictive maintenance adopts a proactive stance. By conducting periodic inspections and utilising sensor data for early fault detection, PdM minimises downtime and maximises equipment efficiency.

3.3 Strategies for Predictive Maintenance Implementation

We explore 2 main approaches. First, we have Anomaly Detection [10]. By analysing various data sources, including sensor readings and historical data, anomalies indicative of potential failures can be identified. This approach allows for the detection of deviations from normal behaviour, prompting timely maintenance actions. Another approach is Prognostics and Health Management (PHM) [11]. Through the development of predictive and degradation models based on collected data and historical trends, PHM enables the forecasting of equipment failures. By predicting potential downtimes, organisations can proactively plan maintenance activities, optimising resource allocation and minimising disruptions.

3.4 Insights from Industry Practices

Numerous organisations have successfully implemented predictive maintenance strategies, realising significant benefits in terms of operational efficiency and cost savings. Case studies across various industries, including manufacturing, aviation, and energy, highlight the effectiveness of PdM in reducing downtime, extending equipment lifespan, and enhancing overall asset reliability. Significant benefits are reaped in terms of reducing the downtime of equipment. Examples include a downtime of just 2.88% at Pepsico plants, a 60% decline in bearing changes at a Noranda alumina plant and a record 1 million mean kilometres between failure at Singapore SMRT train lines [12].

3.5 Gaps

Ensuring the reliability and compatibility of data sources remains a key challenge, particularly in heterogeneous industrial environments with diverse equipment and systems. It is tough to ensure the quality of data collected and how well they integrate with each other. Data processing and cleaning remains a key strategy in ensuring proper prediction. Developing accurate and scalable predictive models requires addressing issues such as data imbalances, model interpretability, and algorithmic complexity. Hence, the algorithm used needs to be robust to ensure accurate prediction. Lastly, organisations must carefully evaluate the cost-benefit implications of implementing predictive maintenance, considering factors such as initial investments in sensor infrastructure, data analytics tools, and workforce training.

4 Exploratory & Data Analysis

Link to all cleaned data and data dictionary:
https://drive.google.com/drive/folders/17SxIGiKkWUCgx2N26rjl4_veO65I0-C2?usp=sharing
Public link to raw data:
https://www.kaggle.com/datasets/behrad3d/nasa-cmaps

4.1 Modelling of run-to-failure dataset

The development of the run-to-failure dataset requires a system model to track and predict the process of degradation of the aircraft engine. In this paper, our run-to-failure dataset uses the C-MAPSS (Commercial Modular Aero-Propulsion System Simulation) as the system model.

The damage modelling of the run-to-failure dataset relies on the application scenario developed which encompasses the assumptions made of the real-world operationality of aircraft engines. The same type of engine will be used in multiple aircrafts operating under various flight conditions. Thus, there are a number of factors contributing to the degree and rate of degradation of the engine after each flight. The application scenario assumes that the degree of degradation after the completion of a flight cannot be estimated solely based on the duration and condition of the flight. Therefore, sensors on the engines are used to collect data during each flight and the data extracted from these sensors will better reflect the degradation caused by each flight.

The scenario of engine degradation is modelled based on two key aspects. Firstly, the degradation of the engine's performance as a result of wear and tear is based on the engine's usage pattern. A new engine will also be subjected to initial wear and tear from manufacturing processes and transportation, which is a reflection of the production of engines in the real-world. The initial wear and tear can result in a significant difference in the total useful operational life of the engine. Secondly, the scenario application also accounts for the impact of maintenance work done on the engine between flights, by incorporating it as process noise. Apart from process noise, there are also other noises that are accounted for, such as the noise that stems from manufacturing processes and discrepancies in assembling of the engines, and measurement noise

4.2 Introduction to CMAPSS Jet Engine Simulated Data

The dataset at the core of our analysis originates from the Commercial Modular Aero-Propulsion System Simulation (C-MAPSS). As originally described in [13] developed under the auspices of NASA's Prognostics Center of Excellence, it was first used as challenge data for the Prognostics and Health Management (PHM) data competition at PHM '08. This dataset has been instrumental in the field of PHM, specifically targeting the predictive maintenance of aircraft turbofan engines. The creation and provision of this dataset stemmed from a critical need in the

prognostics domain: the scarcity of run-to-failure data. Traditional datasets often include data on emerging faults but lack comprehensive coverage of the fault progression up to the point of failure. The C-MAPSS dataset fills this gap by simulating the degradation of aircraft engine modules over time under various operational conditions and fault modes.

The dataset consists of multivariate time series data representing different operational cycles of a fleet of engines, each with its initial wear and manufacturing variations. These engines, simulated under a set of predefined conditions and fault modes, provide a rich source of data for exploring the intricacies of engine degradation. The operational settings included in the dataset—three in number—significantly affect engine performance and are critical to the analysis. Additionally, sensor noise contaminates the data, adding a layer of realism by mimicking real-world operational challenges.

The dataset is structured across four subsets (FD001 to FD004), each simulating different conditions and fault modes. This diversification allows for a comprehensive analysis of engine degradation patterns under a variety of scenarios, providing valuable insights into the factors that influence engine lifespan and reliability.

4.3 Prediction Variable: Remaining Useful Life (RUL)

Central to the analysis is the prediction of the Remaining Useful Life (RUL) of each engine in the test dataset, defined as the number of remaining operational cycles before failure. The objective is to predict the number of operational cycles an engine can run before failure occurs, based on the observed data. By having a good estimate of when an engine fails, decision makers can discern the best time to pause production and conduct an inspection or preventive maintenance, avoiding eventual catastrophic failure and minimising downtime.
We look at the first subset of data for easier visualisation.

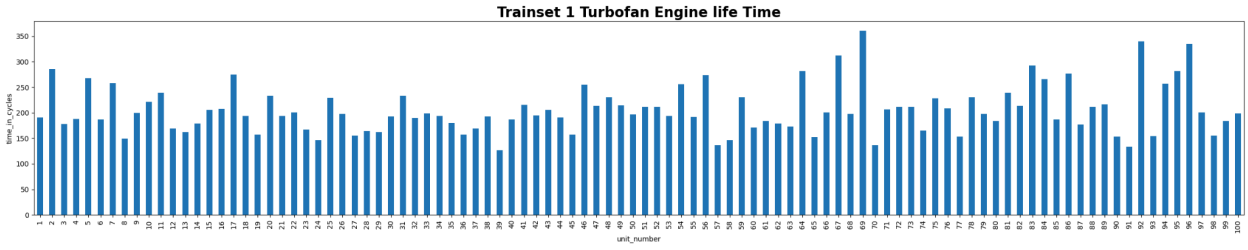


Figure 4.1

As seen in Figure 4.1, the train set for FD001 provides us with 100 engines which we observe throughout their operational life all the way to system failure. While the fleet of 100 engines are of the same type and operate under the same conditions (Sea Level for FD001), each engine starts with different degrees of wear and natural manufacturing variation unknown to the user.

Thus, the total observed engine life amongst the 100 engines have a natural variation around the mean of 205 cycles.

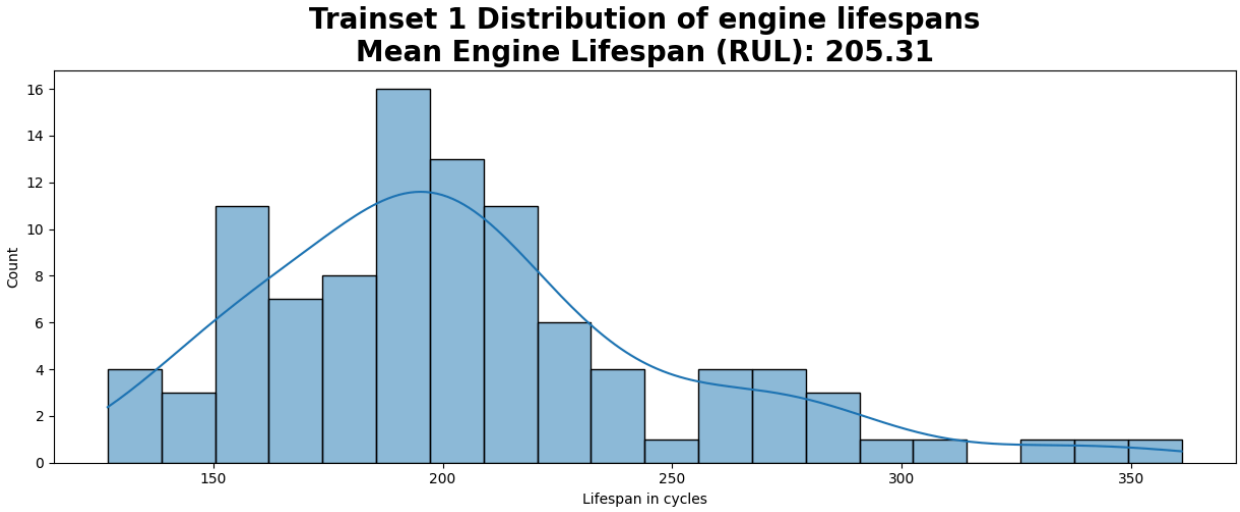


Figure 4.2

As seen in Figure 4.2, The distribution of engine life spans in the train set for FD001 appears right-skewed, meaning there are a number of engines that have a lifespan longer than the mode. Most engines in the fleet have life spans between approximately 150 and 250 cycles, however, the skewness illustrates that there are a few outliers that significantly exceed the average lifespan. Such a distribution is typical in reliability and life data analysis where the item being tested (in this case, engines) is subject to various failure mechanisms that can result in a wide range of life spans.

The test data provides observational data on the lifespans of a similar fleet of engines operating under the same conditions but ends some time prior to system failure. In Figure 4.3, we visualise the operational lives of the fleet of 100 engines in the test set that we can observe as a proportion of their true full lifespans. As we can see, the test set gives a challenging variety of scenarios where some engines fail soon after we observe them while others fail long after we observe them.

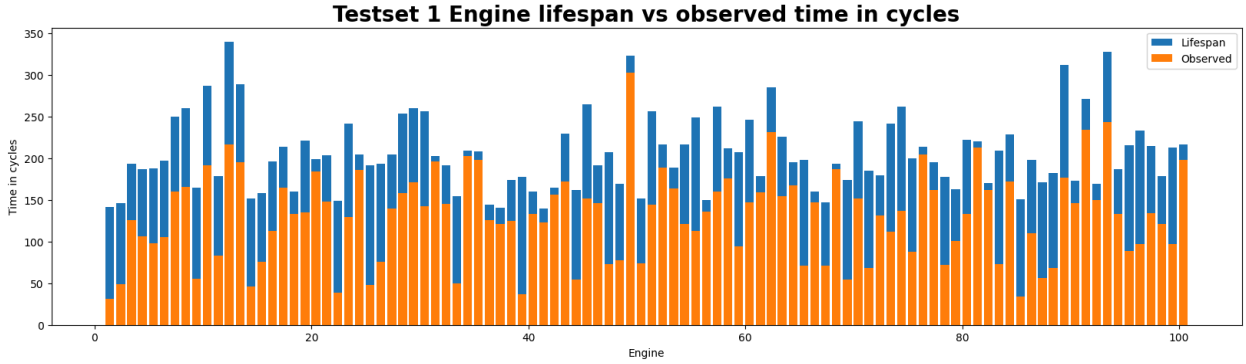
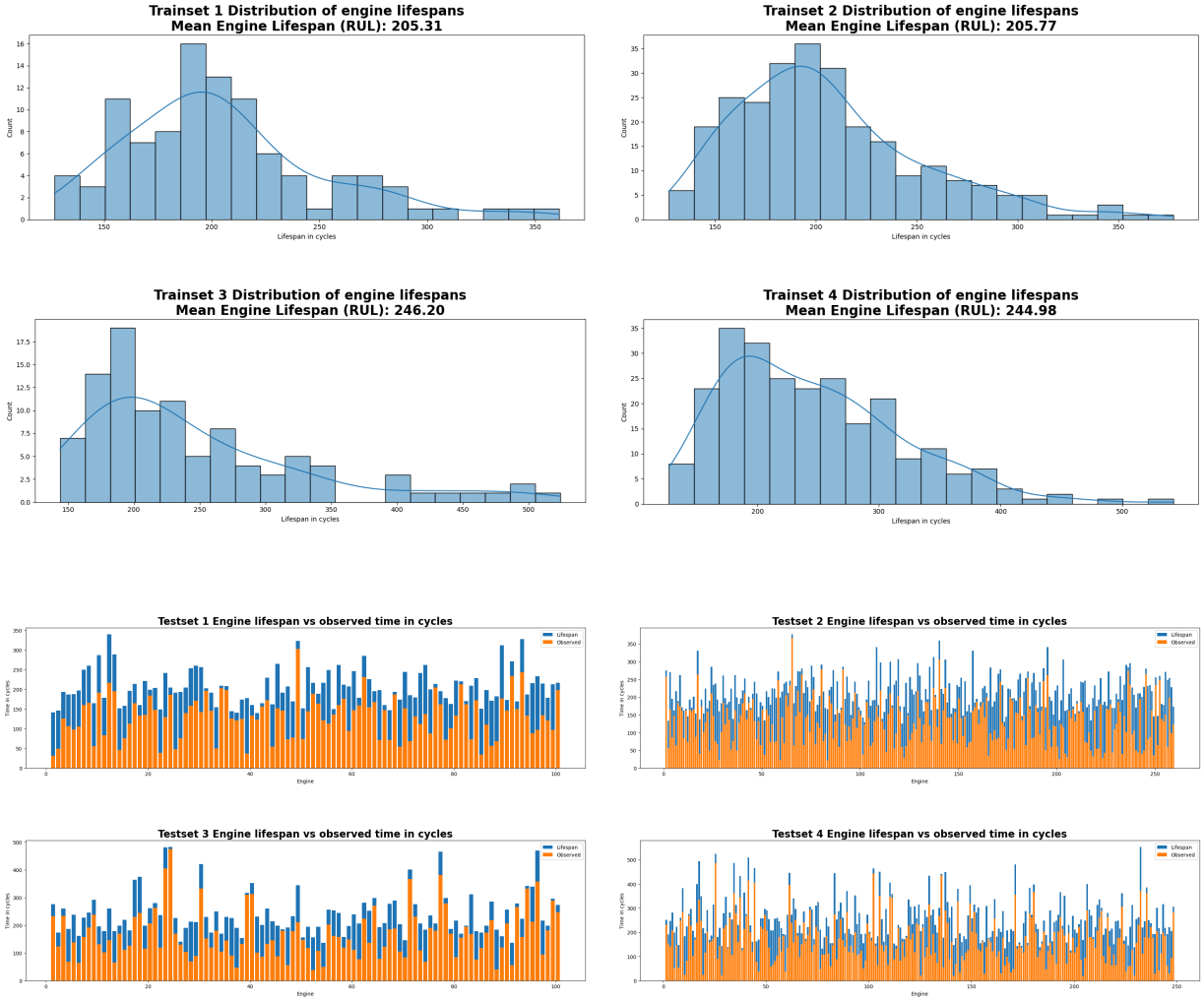


Figure 4.3

The visualisations and distributions of RUL data from all 4 subsets (FD001 to FD004) are provided below.



4.4 Settings and Sensor Data

In addition to RUL data, the dataset includes a slew of features about each engine throughout its lifespan. Presented as individual time series, each engine has 3 operational settings and readings from 21 different sensors. Thus, each data point includes information about the engine's settings and sensor measurements that capture a snapshot of the engine's performance during a single operational cycle. Engine degradation is simulated through changes in operational settings and sensor readings, reflecting the impact of wear and tear as well. This temporal information constitutes the independent variables in our analysis which play a pivotal role in predicting the RUL of each engine.

For visualisation purposes, we plot the time series of features of the engine #1 from the train set of FD001 in Figure 4.4. The x-axis represents the number of operational cycles and the y-axis represents the value of the corresponding setting or sensor measurement.

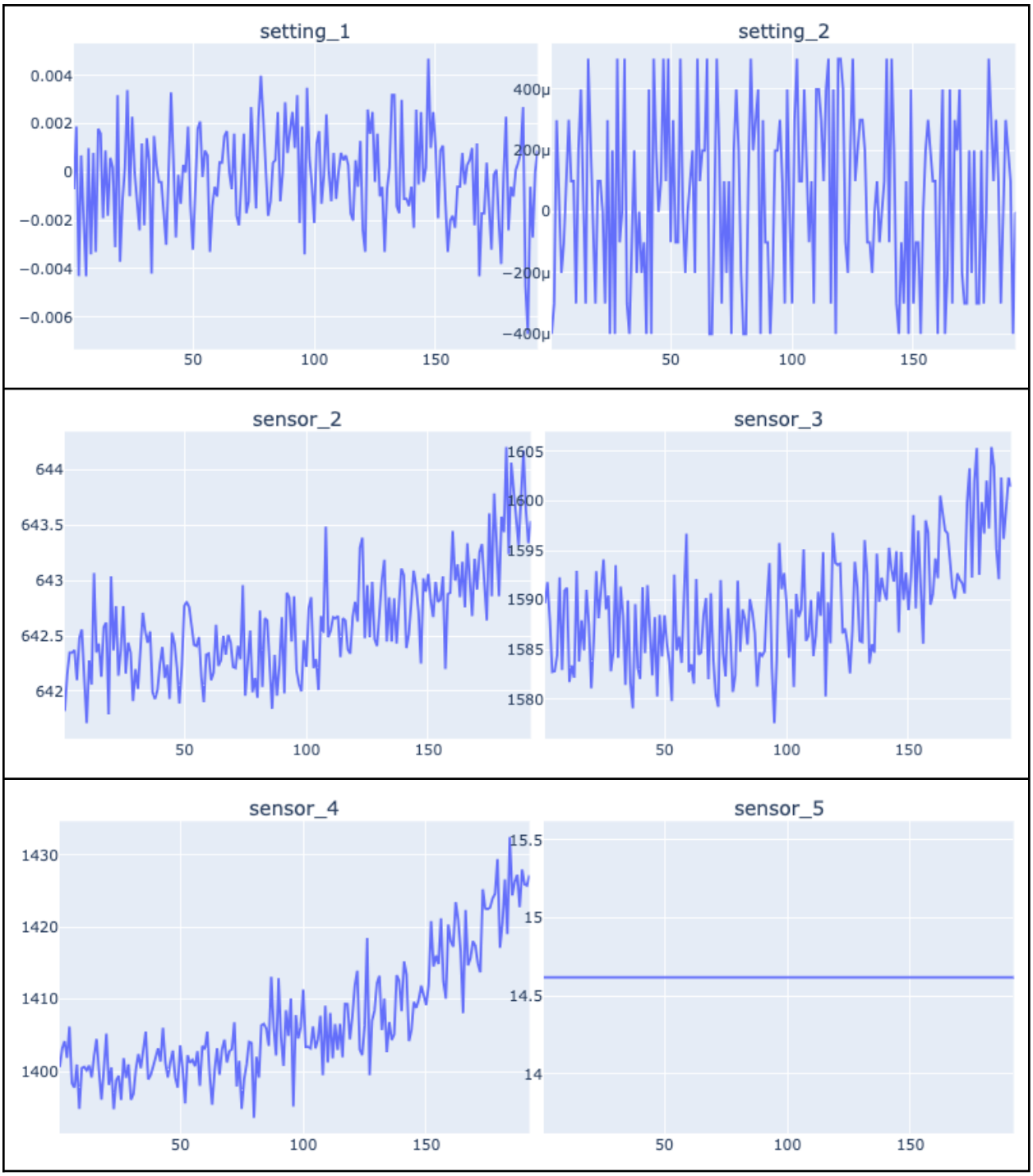


Figure 4.4

The operational settings for engine 1 are changed between a fixed range of values as expected of setting values. For readings of sensor 2, 3, ad 4 of engine #1, we see that the values vary around a mean and eventually trend exponentially higher as the engine reaches system failure. This type of sensor reading is typical of what one might expect from machine failure data. For example, a sensor measuring the temperature of an engine might seem stable at first but increase exponentially as engine functions decay and operate more inefficiently, producing more heat than it is able to handle.

An important point to note is that the reading of sensor 5 remains constant throughout the whole operational life of engine #1. This means that sensor 5's readings are irrelevant to our endeavour to predict RUL and can be removed. Similarly, other sensor measurements contain noisy data and would have to be removed as well. This will be addressed by feature engineering and feature selection methods.

5 Methodology

5.1 Feature Engineering and Feature Selection

5.1.1 Feature engineering

Feature engineering is the process of transforming raw data into informative features that can be used to improve the performance of predictive models. For this project, several steps were undertaken to enhance the dataset for both regression and classification models.

5.1.1.1 Autocorrelation and Stationarity Assessment

Using the Augmented Dickey-Fuller (ADF) test, we evaluated the stationarity of each time series feature within the dataset. Stationarity is a critical characteristic of time series data that ensures the statistical properties do not vary with time. This would be of paramount importance should we choose to use time series specific models to forecast RUL as this would avoid spurious results.

Regardless, our findings from the autocorrelation functions (ACFs) shows the existence of autocorrelation, which is the correlation between a time series and lagged versions of itself. For example in Figure 5.1, we visualise the autocorrelation function of measurements for sensor 7, 8, and 9 of engine #1 from the train set of FD001 and observe that they are highly correlated as evident from the significant ACF values above the shaded blue threshold levels.

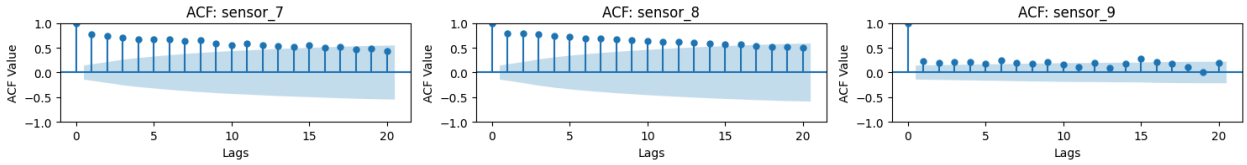


Figure 5.1

5.1.1.3 Temporal Feature Creation

From our findings on autocorrelation above, it is evident that temporal effects are present. Thus, we chose to engineer historical time series features such as moving averages to incorporate these temporal effects and smooth out noisy data, in hopes of capturing the trends and deviations of our time series. These features account for changes in engine behaviour and are expected to enhance the predictive model's performance.

5.1.1.4 Irrelevant Feature Elimination

To streamline the dataset, we removed any features with constant values across all observations since they do not contribute to model differentiation or predictive power.

5.1.1.5 Feature Scaling

Since our setting and sensor feature values differ in magnitude, we have to scale them to prevent our models from having a bias towards features with higher magnitude, potentially leading to incorrect or unstable predictions. We also want to be careful to avoid look-ahead bias, which is when a model inadvertently incorporates data from the future – information that would not be available at the time of prediction – in the training process. Therefore, we chose to standardise our features' time series with a rolling z-score.

5.1.1.6 Classification Labels

To transform our predictive maintenance task into a classification problem, we categorised engine data points into 'at risk' and 'no risk' classes. We make the assumption that an engine was deemed 'at risk' if its RUL was 60 cycles or less (RUL <= 60), while 'no risk' was assigned to engines with an RUL greater than 60 cycles (RUL > 60). This binary classification model aligns with the preventive maintenance objectives, allowing us to target maintenance interventions more effectively.

5.1.2 Feature Selection

Feature selection is the process of identifying and utilising only those features that are most relevant to the predictive task, eliminating redundant or unimportant variables. We study feature importances by applying Least Absolute Shrinkage and Selection Operator (LASSO) and

Random Forest regression to the dataset. LASSO imposes a regularisation term, or constraint that causes regression coefficients for some features to shrink towards zero, eliminating less important features. Random Forest allows us to identify features that are most responsible for decreasing Gini Impurity or variance, and can help us identify features that are most important in explaining the target variable.

We use our findings from both LASSO and Random Forests (Figure 5.2) to determine which features we should use in building our models. For example, we observe that for the train set of FD001, sensor_11_ma_10 and sensor_12_ma_10 seem to be the most important features, which implies that a smoothed time series of sensor 11 and sensor 12 readings could be most indicative for predicting RUL. LASSO coefficients and Random Forest feature importances plots for other train sets can be found in the appendix.

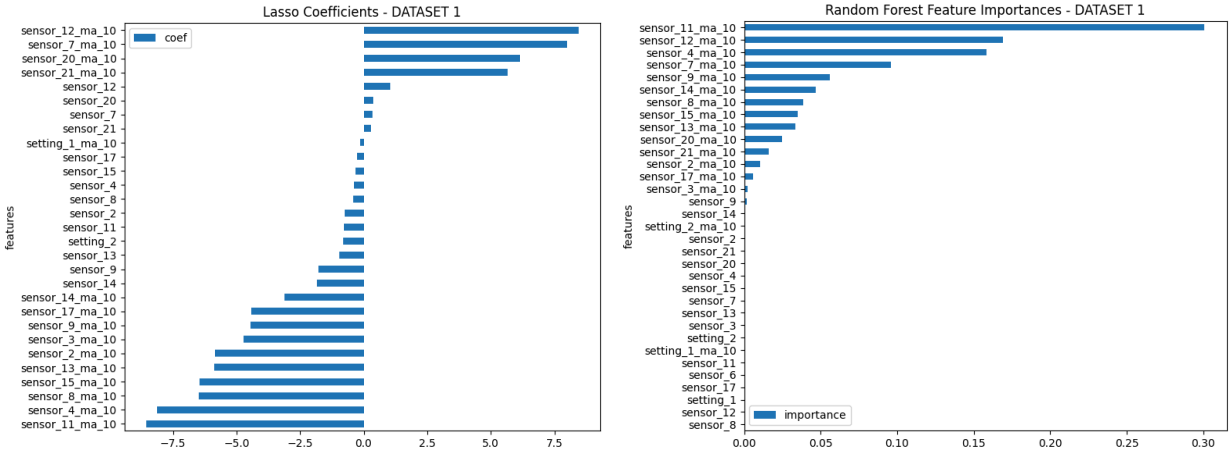


Figure 5.2

5.2 Modelling

We model this problem with both regression and classification models. Traditionally, we can directly predict the value of RUL with regression models. However, as mentioned in section 5.2, we can also transform the prediction of RUL into a classification problem by classifying between classes of engine failure risk (risky and not risky). Regression and classification models have complementary strengths and provide versatile solutions to different risk profiles and scenarios. For instance, regression models can provide a direct prediction of RUL, which may be valuable for precise scheduling of maintenance. On the other hand, classification models can categorise the state of equipment into broader categories such as ‘at risk’ or ‘not at risk’, which can be useful for prioritising maintenance actions when dealing with several components. Furthermore, different maintenance tasks and decisions might have different timelines and risk profiles. Classification models can be tuned to be more conservative and prioritise minimising either false

positives or false negatives by setting different thresholds. Conversely, regression models offer a more nuanced view of the risk based on the predicted RUL value. We find value in exploring both problem sets and have employed a diverse range of machine learning models to tackle both scenarios.

Models used in predicting RUL - regression problem:

1. Linear Regression: Simple and intuitive interpretation for maintenance planning. However, it is not without its drawbacks as the relationships present in the data can be seen to be non-linear.
2. Elastic Net Regression: Combines L1 and L2 regularisation techniques to reduce variance in our forecasts, ensuring a more robust model than a simple linear regression.
3. Random Forest Regressor: Offers high practical performance across a wide range of industries and use-cases especially for complex and large datasets.
4. XGBoost Regressor: Heralded as the industry standard as it is able to capture non-linear patterns efficiently while also being scalable.
5. Support Vector Regression: Effective in datasets with many independent variables and capable of capturing non-linear relationships.

Models used in predicting risk classes - classification problem:

1. Gaussian Naive Bayes: A quick and simple model suitable for initial baseline models.
2. Decision Tree: Intuitive model structure that is easy to interpret. It is useful for understanding key factors influencing risk.
3. Random Forest Classifier: High practical performance, similar to its regression counterpart, and is less prone to overfitting.
4. XGBoost Classifier: Boosted tree methods are proven to excel in performance across many classification problems.
5. K-Nearest Neighbors: A non-parametric method that classifies based on the majority vote of neighbours, without relying on learning a pre-defined relationship.

This initial exploration allows us to understand the baseline performance of each model. Once we have identified the most promising models, we can then iteratively test different combinations of hyperparameters to enhance the model's ability to generalise unseen data, and hopefully improve predictions of RUL.

6 Model Results and Discussion

In this section, we present the results from our machine learning models and discuss some of the potential implications on Aramco's maintenance procedures. For the sake of brevity, our analysis

will only be presented using data from train set FD001, however results for other datasets can be found in the appendix for reference.

6.1 Regression Models Results

We evaluate the efficacy of our regression models in predicting RUL using 4 key metrics:

1. Mean Absolute Error (MAE): MAE describes the average errors in a set of predictions, without considering the direction of errors. It is a very interpretable metric since it retains the units of RUL. Since MAE gives linear penalties to our errors, it is not as sensitive to outliers. At the same time, this may not reflect the severity of the worst errors.
2. Mean Squared Error (MSE): MSE punishes larger errors quadratically since it calculates squared errors. A high MSE and low MAE could therefore point to outliers present in the predictions. MSE is emphasised as a main evaluation metric if larger errors are particularly undesirable.
3. Root Mean Squared Error (RMSE): RMSE is the square root of MSE, thereby providing error magnitudes in the same units as our original data, making it more interpretable.
4. Coefficient of Determination (R^2): R^2 describes the goodness-of-fit of our model and is derived by comparing the sum of squared residuals and the variance of our response variable. A higher R^2 indicates our model can explain most of the variability of the response date around its mean. However, R^2, like other metrics above, cannot determine if there is an inherent bias in our model and for that we have to reference residual plots.

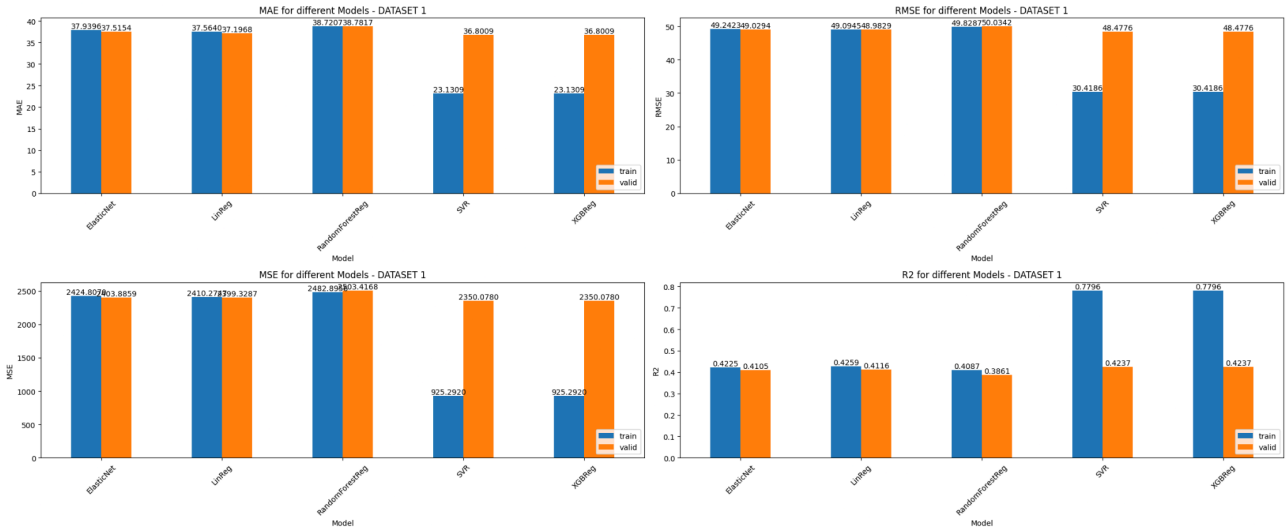


Figure 6.1

Looking at figure 6.1, we observe that the performance of regression models is generally subpar with less than 0.5 R^2 values. However, XGBoost stands out as a notable model across all metrics between training and validation sets. It registered the lowest RMSE and highest R^2 values. However, we note that it still fails to generalise for unseen data given that its validation set

metrics are far worse than that of its training set. Despite this, given XGBoost's performance and overall ability to capture complex non-linear relationships, we chose it to be the model we would optimise via hyperparameter tuning with grid search.

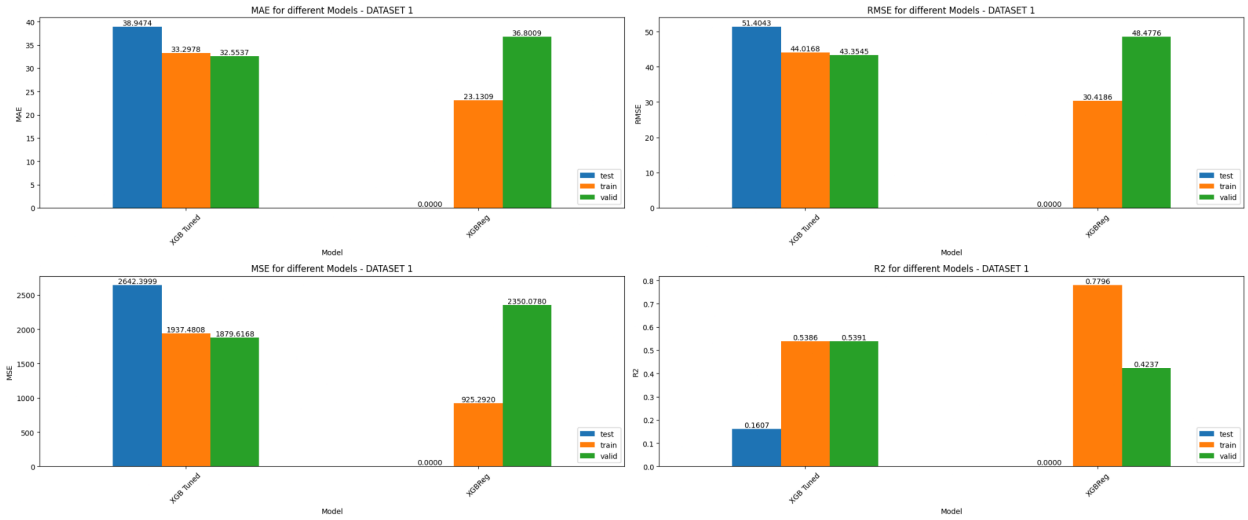


Figure 6.2

Post hyperparameter tuning, we find that the tuned XGBoost model seems to perform worse than the baseline model in all metrics for the train set in Figure 6.2. However, it can also be seen that the tuned model has better validation metrics, implying that it generalises better for unseen data. While this may be critical for real-world applications, testing the model on our test set shows otherwise.

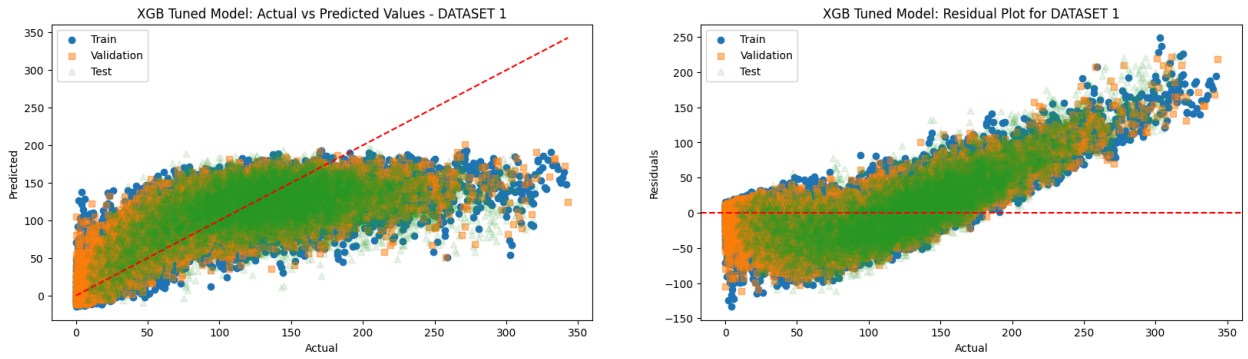


Figure 6.3

Figure 6.3 also shows the inherent bias our model predictions have. It consistently underestimates the actual RUL for high RUL values and overestimates for low RUL values. The inherent bias and variance in this model shows that it is not suitable to be deployed in real world settings, despite the advantage it offers in directly predicting RUL.

6.2 Classification Models Results

For our classification models, the binary labels defined in section 5.2.1.6 is based on the assumption that an RUL of 60 implies that predictive maintenance should be scheduled soon. While the threshold of 60 RUL is defined heuristically, it is important to note that it can be adjusted contingent on expert domain knowledge and strategy adaptations. For example, it can be decreased for a more granular view on engines to schedule for maintenance or increased to be more conservative.

When evaluating our classification models, we utilise 4 key metrics to gauge their ability to predict the maintenance needs:

1. Area Under the Receiver Operating Characteristic Curve (ROC AUC): ROC AUC measures the model's ability to distinguish between 'at risk' and 'not at risk' engines. Intuitively, a ROC AUC of 0.9 refers to the probability the model will output a higher value for a randomly chosen positive class than a randomly chosen negative class. A high ROC AUC implies good separability between the distributions of classes. The ROC is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings, which is the cutoff value for classifying observations into different classes. Note that classification models often output a probability or a score that indicates the likelihood of an observation belonging to the positive class, which in this case is 'at risk'. The cutoff is also usually set at 0.5. By having the flexibility to define different threshold settings, companies can alter their models to prioritise minimising TPR or FPR based on their needs.
2. Recall: Recall or Sensitivity describes the model's coverage of actual positive or 'at risk' samples. This is critical when the cost of false negatives is very high (false negatives), which could likely be the case for predictive maintenance as it potentially leads to unplanned downtime.
3. Accuracy: Accuracy reflects the overall performance of the model and reflects the ratio of the total predictions that the model correctly classifies. However, given our unbalanced dataset where 'not at risk' cases may significantly outnumber 'at risk' cases, accuracy can be misleading. It might not adequately represent the model's performance on the minority class, which is often of more interest in predictive maintenance.
4. Precision: Precision measures how accurate the 'at risk' classifications are. In contexts where the cost of false positives is high, precision becomes a key metric to optimise. In the case of predictive maintenance, Precision however, may not be as important since it is safer to wrongly classify an engine that is 'not at risk' (as 'at risk') than to wrongly classify an engine that is 'at risk' (as 'not at risk').

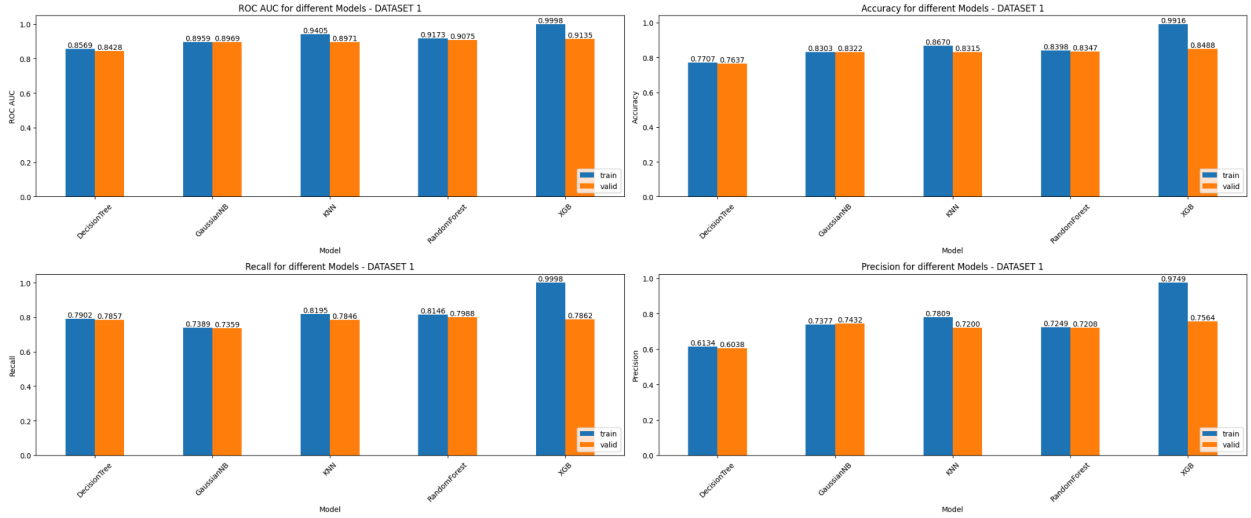


Figure 6.3

In examining the performance of classification models as depicted in Figure 6.3, we observe that XGBoost outperforms other models – its highest ROC AUC score points towards its excellent ability to distinguish between classes. It is also worth noting that almost all our classification models have good performance metrics – they generally have high ROC AUC and Recall scores across all models, especially in Random Forest and XGBoost. This signifies their ability to distinguish between classes and accurately predict whether an engine is ‘at risk’ or ‘not at risk’. This is in stark contrast to the poor performance metrics we see in our regression models.

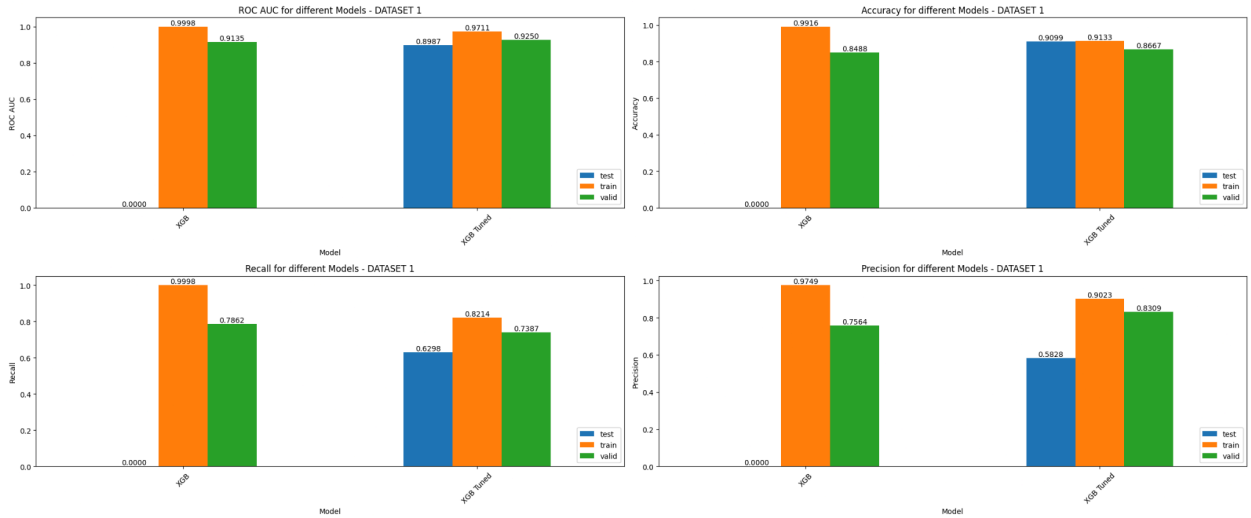


Figure 6.4

We choose to optimise the XGBoost model’s hyperparameters and it has led to improvements in the model’s ability to generalise unseen data. While there is a slight decrease in performance metrics post-tuning as seen in Figure 6.4, the decline is marginal and acceptable given the model’s improved generalisation capabilities which signifies little overfitting. In fact, this slight

trade off in raw performance is often an indicative sign of a model that is less likely to overfit and thus more robust against the variability inherent in real-world data.

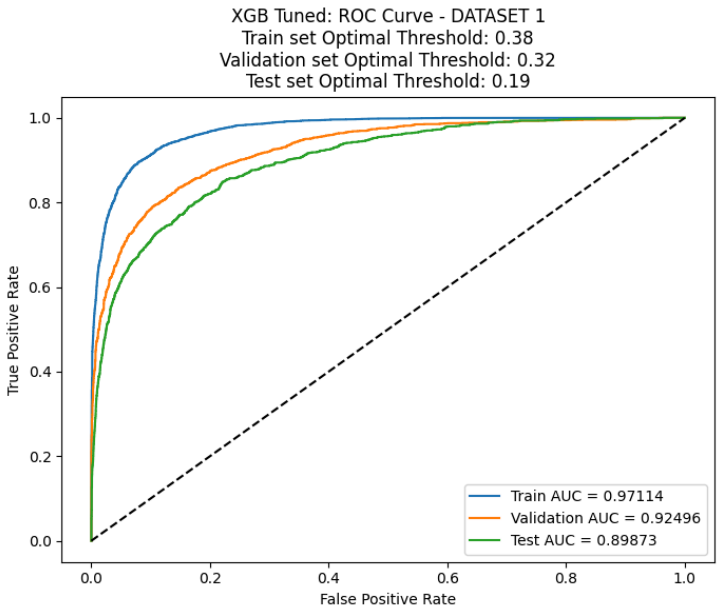


Figure 6.5

Figure 6.5 shows the consistency in the XGBoost classifier's ability to distinguish between classes. The consistency in validation and testing performance across the board suggests that the tuned XGBoost classifier can be expected to perform with similar efficacy in practical real-world scenarios, allowing us to accurately predict 'at risk' and 'not at risk' engines.

6.3 Discussion

Given that the classification algorithms showed better results on test data in accurately predicting when an engine is 'at risk', it is clear that they can serve as a powerful tool for prioritising maintenance scheduling tasks. Unlike regression models, classification models directly provide a binary output that corresponds to actionable categories, on the account that RUL cutoffs have been defined by a subject matter expert.

Furthermore, classification models facilitate a more nuanced cost-benefit analysis by computing the expected profit or loss across different models and thresholds. We use the methodology described in [6] to calculate the expected monetary value by considering the costs and benefits associated with true positives, true negatives, false positives, and false negatives. It assigns concrete costs and benefits to each outcome and is computed in line with the confusion matrix. The expected profit (or loss) is calculated as follows:

$$Expected\ Profit\ =\ P_P[TPR * B_{TP}\ +\ FNR * C_{FN}]\ + P_N[\ TNR * B_{TN}\ +\ FPR * C_{FP}]$$

where P refers to the proportion of positive or negative class instances in the test data. In essence, this provides a clear method to quantify the trade-offs between different types of errors. We assume $B_{TP} = 400k$, $B_{TN} = 0$, $C_{FN} = -500k$, $C_{FP} = -40k$. Note that these values have to be defined by someone with subject matter expertise, but for the purposes of this project, we assume that it costs significantly more to misclassify an engine 'at risk' as 'not at risk' since this implies the engine would eventually fail and costs would encompass unexpected downtime, additional repair costs and potential secondary damage.

Recall that the number of true positives, false negatives, true negatives and false positives changes with different thresholds as discussed in section 6.2. Therefore, using different thresholds also enable us to derive different expected profit calculations. This allows companies to set different parameters based on their risk profile and overall goal. Looking at Figure 6.6, we see that different thresholds lead to different TPR, FNR, TNR and FPR which yield different expected profits. Observe that the optimal threshold is found to be 0.082. This low threshold also implies that the model is placing more emphasis on minimising the number of false negatives.

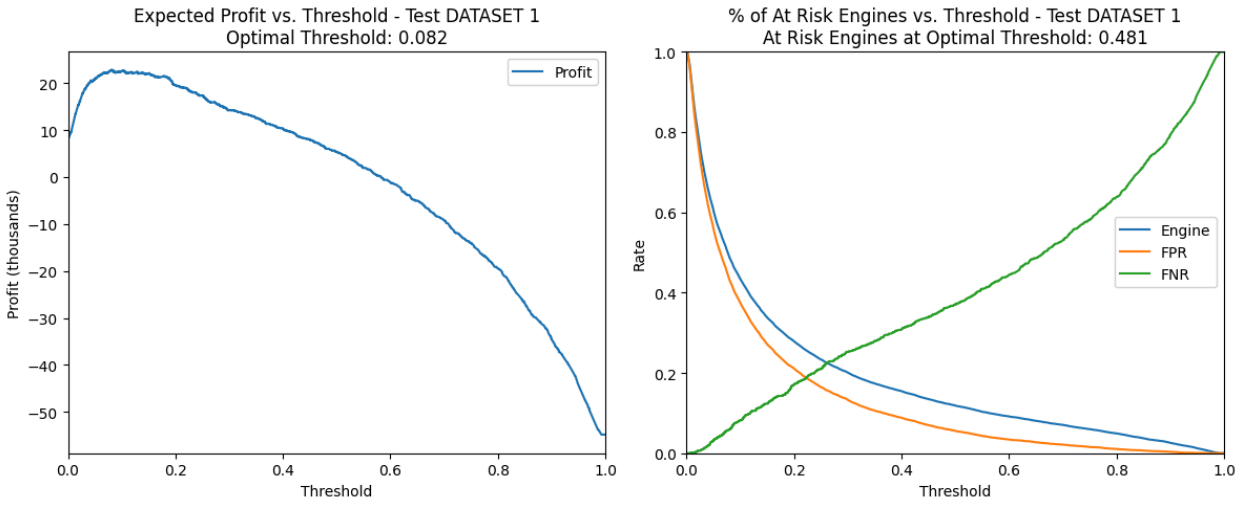


Figure 6.6

As a result of varying thresholds, the number of engines classified as 'at risk', seen in Figure 6.6, will also change and by extension the amount of resources a company has to set aside for maintenance scheduling. This can also serve as another important factor for companies to schedule and plan so as to ensure that the selected model and threshold settings do not exceed operational or logistical capacities.

In practice, this can translate into a tangible advantage for maintenance scheduling and resource allocation. Using the threshold that optimises the Expected Value, we can target maintenance efforts where they are most economically justified, potentially reducing unnecessary downtime and focusing resources on preventing costly failures. This method aligns with proactive maintenance strategies, avoiding the pitfalls of both reactive and excessively conservative approaches.

In summary, the enhanced capability of classification models to incorporate economic reasoning into predictive analytics offers a compelling reason to favour them over regression models in certain contexts of predictive maintenance, particularly when paired with the superior predictive performance evidenced in our tests.

7 Recommendations

Aramco can utilise both XGBoost regressor and XGBoost classifier to automate the maintenance of their key equipment, reducing resources required for such purposes. Firstly, for XGBoost regressor, the regression model provides an ideal solution for equipment that requires precise scheduling of maintenance due to the greater care they require or their more expensive nature. Aramco can apply this to the following equipments:

1. Gas turbines and compressors, which are critical components in oil and gas production.
2. Drilling equipment, such as drilling rigs and pumps.
3. Pipeline monitoring systems which are essential for transporting oil and gas over long distances.

These are critical and expensive components for Aramco and maintaining their optimal performance is crucial for efficiency. Therefore, by utilising XGBoost regressor to predict RUL and schedule precise maintenance, Aramco will be able to maximise their usage at their optimal stage while avoiding the risk of failures. Secondly, Aramco can use XGBoost classifier for the maintenance of less critical equipment, or for maintenance actions that have multiple components, by classifying them as ‘at risk’ or ‘not at risk’, prioritising what actions they should take first based on general thresholds rather than precise RUL predictions. This is especially important to help Aramco avoid periods of downtimes for several equipment and to help them better handle their capacity for maintenance to avoid situations where there is a lack of resources or manpower for maintenance. Aramco can apply this to the following equipments:

1. Light vehicles, generators, and non-critical machinery.
2. Facilities-related equipment like HVAC systems.
3. Systems supporting operations such as IT infrastructure and communication systems.

Furthermore, the classifier can help Aramco minimise false positives, which are unnecessary maintenance to reduce cost and utilisation of resources, and false negatives, which are missed maintenance resulting in downtime and greater loss for the company as they need to replace the

entire equipment and have failed to meet their demand. To maintain efforts in implementing data-driven solutions, Aramco should also conduct comprehensive training programs for their middle and senior management to educate them and promote a data-driven culture in the company.

8 Conclusion

In conclusion, this project has demonstrated the significant value that data analytics, particularly through the utilisation of machine learning models, can offer to Aramco in enhancing the efficiency and reliability of its maintenance operations. By leveraging both XGBoost regressor and classifier, Aramco can automate maintenance scheduling for a wide range of equipment, from critical components like gas turbines and compressors to less critical machinery and facilities-related equipment.

The XGBoost regressor provides precise predictions of RUL, enabling Aramco to schedule maintenance with pinpoint accuracy for critical equipment, thereby maximising uptime and minimising the risk of costly failures. On the other hand, the XGBoost classifier offers a binary classification of 'at risk' or 'not at risk', facilitating prioritisation of maintenance actions for less critical equipment and allowing for efficient resource allocation.

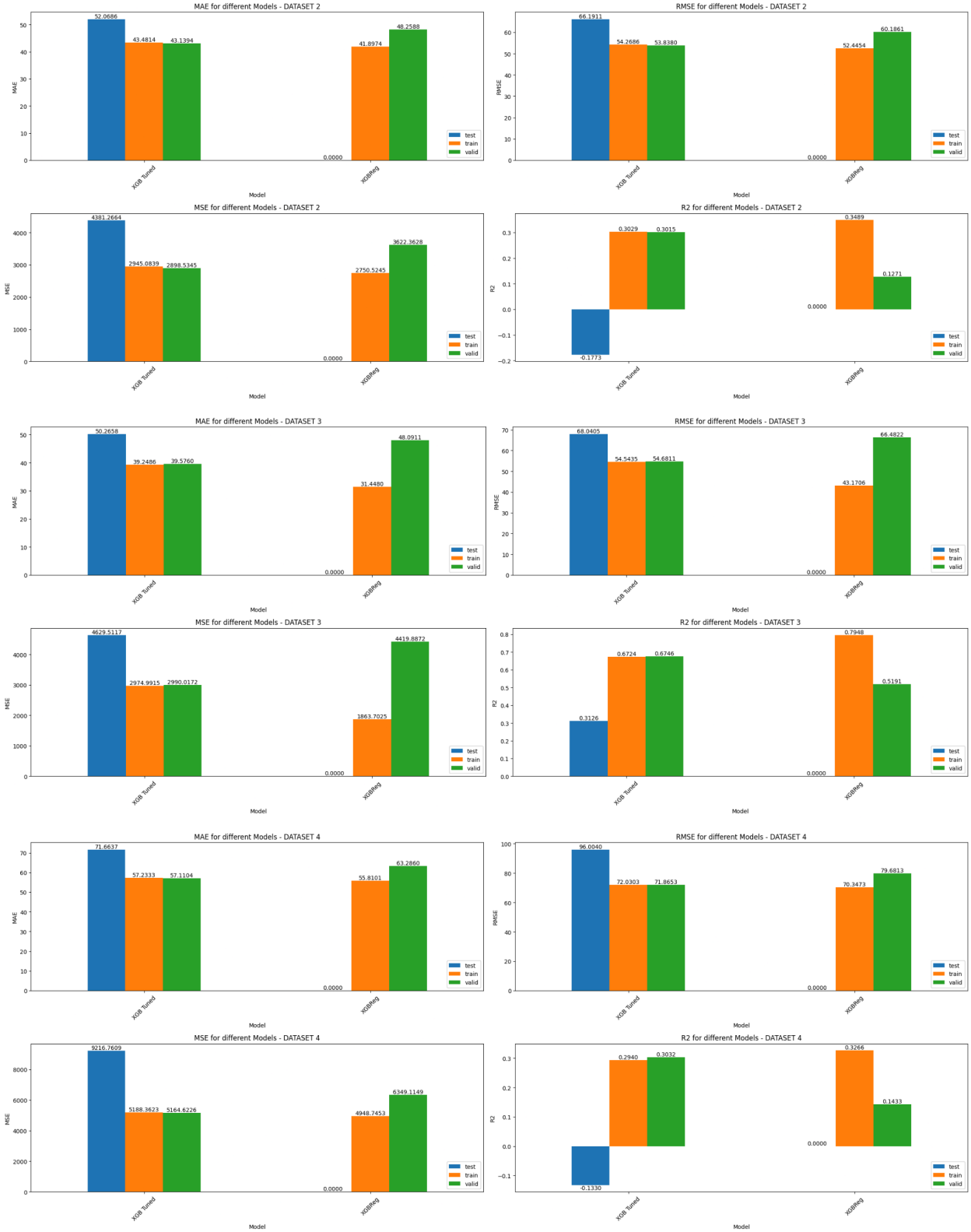
Moreover, the incorporation of economic reasoning into predictive analytics, as demonstrated through the calculation of expected profit and the optimization of thresholds, ensures that maintenance efforts are aligned with Aramco's strategic objectives and risk profile. By minimising both false positives and false negatives, Aramco can optimise resource utilisation and avoid unnecessary downtime, ultimately driving cost savings and operational efficiency.

In time to come, it is clear that data analytics will continue to play a pivotal role in shaping the future of maintenance in the oil and gas industry. As technology advances and data becomes increasingly abundant, Aramco must remain proactive in embracing data-driven solutions and fostering a culture of innovation and continuous improvement. Comprehensive training programs for middle and senior management will be essential in ensuring widespread adoption of data analytics practices across the organisation, such as predictive maintenance, ultimately positioning Aramco for sustained success in a rapidly evolving industry landscape.

Overall, this project highlights the transformative potential of data analytics in driving operational excellence and competitive advantage for Aramco, paving the way for a future where maintenance operations are not just reactive, but proactive and predictive, enabling Aramco to stay ahead of the curve and thrive in an ever-changing market environment.

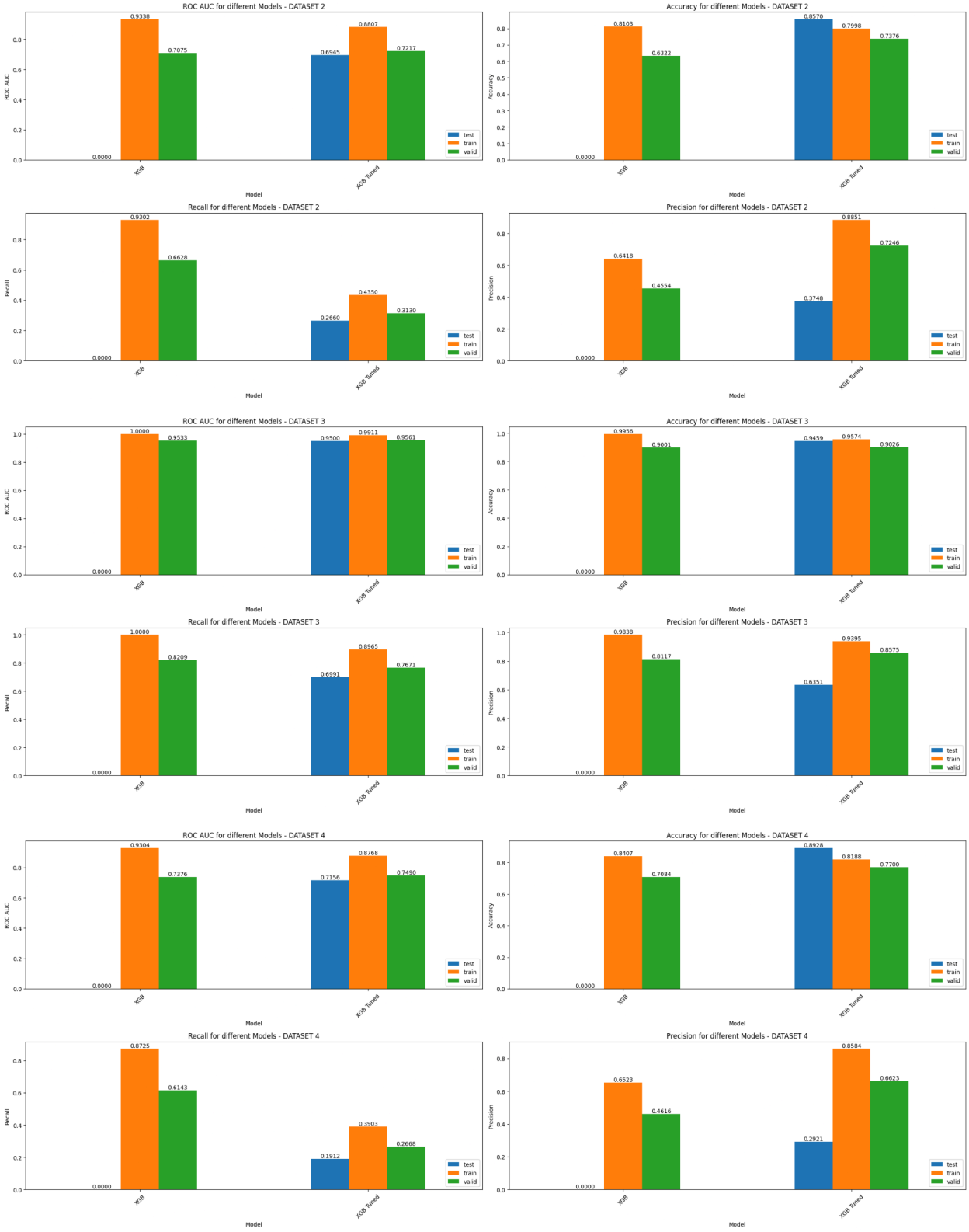
9 Appendices

Appendix Fig A: Regression Model Metrics for Datasets FD0002, FD0003, FD0004

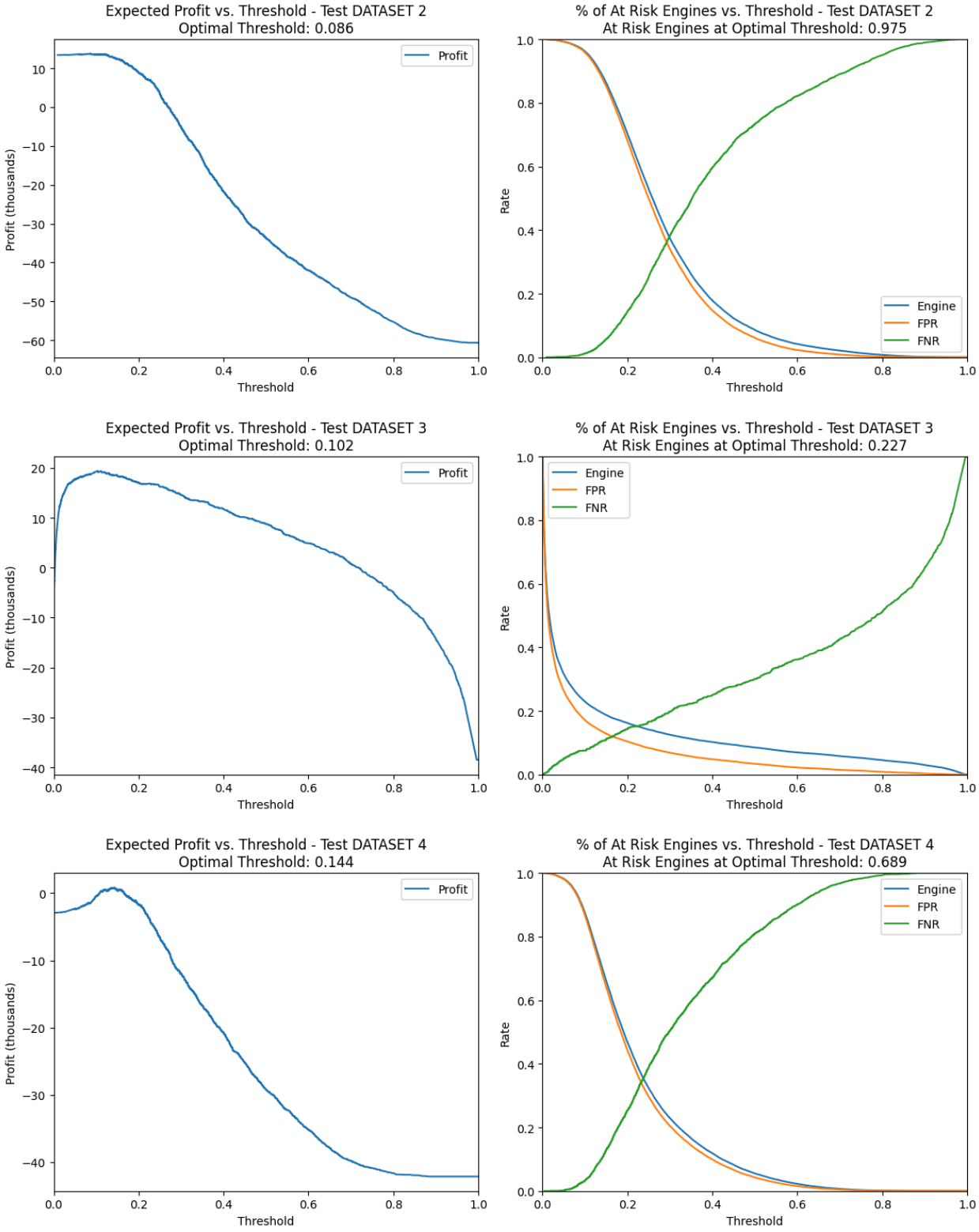


Appendix Fig B: Tuned XGB vs. Baseline XGB Model Metrics for Datasets FD0002, FD0003, FD0004

Appendix Fig C: Classification Model Metrics for Datasets FD0002, FD0003, FD0004



Appendix Fig D: Classification Tuned XGB vs. Baseline XGB Model Metrics for Datasets FD0002, FD0003, FD0004



Appendix Fig E: Expected Profit and % of Engines at Risk across different thresholds for the tuned XGB classifier for different datasets

10 References

[1] “Senseye Report: Downtime is Costly,” Turbomachinery Magazine. Accessed: Apr. 08, 2024. [Online]. Available: https://www.turbomachinerymag.com/view/senseye-report-downtime-is-costly

[2] “Aramco announces full year 2023 results.” Accessed: Apr. 08, 2024. [Online]. Available: https://www.aramco.com/en/news-media/news/2024/aramco-announces-full-year-2023-results

[3] “CMAPSS Jet Engine Simulated Data | NASA Open Data Portal.” Accessed: Apr. 08, 2024. [Online]. Available: https://data.nasa.gov/Aerospace/CMAPSS-Jet-Engine-Simulated-Data/ff5v-kuh6/about_data

[4] “Predictive Maintenance in Oil and Gas Industry,” UptimeAI. Accessed: Apr. 08, 2024. [Online]. Available: https://www.uptimeai.com/resources/predictive-maintenance-in-oil-and-gas-industry/

[5] N. Burclaff, “Research Guides: Oil and Gas Industry: A Research Guide: Upstream: Production and Exploration.” Accessed: Apr. 08, 2024. [Online]. Available: https://guides.loc.gov/oil-and-gas-industry/upstream

[6] N. Ohalete, A. Aderibigbe, E. Ani, P. Ohenhen, and A. Akinoso, “Advancements in predictive maintenance in the oil and gas industry: A review of AI and data science applications,” *World J. Adv. Res. Rev.*, vol. 20, pp. 167–181, Dec. 2023, doi: 10.30574/wjarr.2023.20.3.2432.

[7] “Digital technologies.” Accessed: Apr. 08, 2024. [Online]. Available: https://www.aramco.com/en/what-we-do/energy-innovation/digitalization/digital-technologies

[8] A. Bousdekis, K. Lepenioti, D. Apostolou, and G. Mentzas, “Decision Making in Predictive Maintenance: Literature Review and Research Agenda for Industry 4.0,” *IFAC-Pap.*, vol. 52, no. 13, pp. 607–612, Jan. 2019, doi: 10.1016/j.ifacol.2019.11.226.

[9] “Run-to-Failure Relay Dataset for Predictive Maintenance Research With Machine Learning | Semantic Scholar.” Accessed: Apr. 02, 2024. [Online]. Available: https://www.semanticscholar.org/paper/Run-to-Failure-Relay-Dataset-for-Predictive-With-Winkel-Deuse-Kleinsteuber/1242c47046db7a6d289d9d83bc8dddec0da7e285

[10] J. Carrasco *et al.*, “Anomaly Detection in Predictive Maintenance: A New Evaluation Framework for Temporal Unsupervised Anomaly Detection Algorithms,” *Neurocomputing*, vol. 462, pp. 440–452, Oct. 2021, doi: 10.1016/j.neucom.2021.07.095.

[11] R. Li, W. J. C. Verhagen, and R. Curran, “Toward a methodology of requirements definition for prognostics and health management system to support aircraft predictive maintenance,” *Aerosp. Sci. Technol.*, vol. 102, p. 105877, Jul. 2020, doi: 10.1016/j.ast.2020.105877.

[12] S. Kennedy, “Push the needle: How 6 companies are achieving predictive maintenance success,” Plant Services. Accessed: Apr. 08, 2024. [Online]. Available: https://www.plantservices.com/predictive-maintenance/predictive-maintenance/article/11288555/push-the-needle-how-6-companies-are-achieving-predictive-maintenance-success

[13] A. Saxena, K. Goebel, D. Simon, and N. Eklund, “Damage propagation modeling for aircraft engine run-to-failure simulation,” *Int. Conf. Progn. Health Manag.*, Oct. 2008, doi: 10.1109/PHM.2008.4711414.